Vers une étude comparative diachronique des mondes lexicaux du féminisme

Stéphanie Léon, Mathieu Roche

Vers une étude comparative diachronique des mondes lexicaux du féminisme

Stéphanie Léon*,**, Mathieu Roche**

* Université de Provence, 29 avenue Robert Schuman, 13621 Cedex 1

** LIRMM, CNRS, Université Montpellier 2, France

Résumé. Cet article présente une approche lexicale d'analyse comparative diachronique entre deux corpus traitant du féminisme, sur deux périodes différentes. L'analyse lexicale s'appuie sur la collecte des « mondes lexicaux » (unités lexicales simples et complexes qui sont significativement fréquentes) liés aux deux corpus et sur une analyse comparative de ces mondes lexicaux. Les résultats montrent que les unités lexicales simples sont très proches entre les deux corpus qui traitent de la même thématique, tandis que les unités lexicales complexes sont significativement différentes, car plus spécialisées à une sous-thématique et à une période.

1 Introduction

Le Centre d'Etudes Alexandrines[1] a entrepris un vaste travail de numérisation de la presse francophone d'Egypte, sur une période de deux cents ans, depuis l'importation de la première presse par Bonaparte en juillet 1798. L'objectif est de mettre à la disposition de la communauté des chercheurs les journaux, revues, périodiques francophones publiés sur le sol égyptien, comme par exemple le Courrier et la Décade égyptienne publiés par Bonaparte entre 1798 et 1801 ou encore la Réforme illustrée des années 1950. L'aspect éphémère de la masse de documents contenue dans ces publications entraînait le risque qu'ils soient négligés voire oubliés par les historiens. Pourtant ces documents contiennent des informations au jour le jour sur l'histoire de l'Égypte sous toutes ses facettes.

Déjà plus d'une dizaine de milliers de pages sont disponibles, non pas en mode image, mais en mode de texte intégral (au format PDF). Suite à ce projet, un travail d'analyse lexicale automatisée s'est mis en place, avec pour objectif d'extraire les « mondes lexicaux » de ces données, selon leur thématique, leur période, etc., en vue de permettre des recherches lexicales et des analyses comparatives. L'objectif de nos travaux est d'extraire les principales caractéristiques lexicales de ces revues, en pro-

[1] http://www.cealex.org/

posant une méthodologie réapplicable par la suite sur un grand nombre d'autres données. Pour cette étude, nous nous limitons à deux revues (une revue francophone égyptienne datant respectivement des années 1930 et l'autre, contemporaine et française) et à une thématique, le féminisme.

Le féminisme est un mouvement social qui a pour objet l'émancipation de la femme, l'extension de ses droits en vue d'égaliser son statut avec celui de l'homme, en particulier dans le domaine juridique, politique et économique. Dans un tel contexte, la presse féminine que nous étudions dans cet article a toujours eu un rôle crucial consistant à s'adresser exclusivement aux femmes en traitant de leur problème. Une telle démarche consistant à reconnaître l'identité des femmes est considérée comme éminemment féministe[2]. Notons cependant que certains aspects des combats féministes ne sont pas toujours abordés et revendiqués de manière affirmée dans ce type de presse.

Dans cet article, nous analysons le thème du féminisme d'une part dans une perspective comparative diachronique, et d'autre part, par analyse thématique contrastive, entre le féminisme et la critique littéraire. Nous faisons l'hypothèse qu'une analyse comparative lexicale permet de faire émerger les similarités thématiques entre deux revues, mais aussi les spécificités lexicales liées à une sous-thématique ou à une période donnée.

2 Objectifs

L'intérêt de notre projet est de permettre au chercheur (qu'il soit historien, linguiste, etc.) de pouvoir regrouper automatiquement plusieurs revues en fonction de critères précis, comme par exemple en fonction d'un événement (constituer une sélection d'articles transversaux sur un même événement de l'histoire du pays ou de l'histoire de la presse du pays), ou en fonction d'une thématique (la censure, les campagnes de presse, l'insulte et la diffamation…). Il s'agit également d'avoir accès au monde lexical d'une revue afin de connaître ses unités lexicales significatives.

Les aspects lexicaux que nous faisons émerger sont de deux ordres. Nous distinguons deux phénomènes de représentation du contexte d'un mot ou d'une combinaison lexicale, regroupés sous la notion de « monde lexical». D'une part, nous faisons émerger les « mondes lexicaux » des différents corpus étudiés, par une extraction d'Unités Lexicales Simples, répertoriées par catégorie grammaticale « pertinente ». Ces co-occurrences constituent l'entourage lexical, sans prendre en compte les relations de dépendance syntaxique entre les unités lexicales. Nous faisons l'hypothèse que ces mondes lexicaux, représentatifs d'une thématique donnée et du vocabulaire pertinent, permettent d'extraire des régularités entre plusieurs corpus, mais aussi de faire émerger des différences, qu'elles soient sémantiques ou diachroniques.

[2] Femmes d'Aujourd'hui (Périodique). Presse féminine et féminisme. In: Les Cahiers du GRIF, N. 23-24, 1978. Où en sont les féministes ? pp. 120-123

Dans cette étude, nous avons choisi d'étudier la même thématique, mais à une période différente. Notre comparaison des mondes lexicaux se veut donc diachronique. D'autre part, nous nous intéressons aux patrons morpho-syntaxiques (relations de dépendance syntaxique) les plus récurrents, pour chaque corpus. Dans cet article, nous parlerons d'Unité Lexicale Complexe afin de désigner ces associations lexicales. La section suivante décrit la méthode d'extraction des mondes lexicaux formés des Unités Lexicales Simples et Complexes. La section 4 présente une analyse des mondes lexicaux obtenus. Enfin, quelques perspectives sont présentées en section 5.

3 Extraction des Mondes Lexicaux

3.1 Processus global d'extraction des Mondes Lexicaux

Le processus global proposé est composé de quatre étapes successives. La première étape consiste à acquérir le corpus. Cette phase d'acquisition est détaillée dans la section 4.1 de cet article. La deuxième étape effectue une tâche de normalisation du corpus par un processus de « nettoyage » et d' « uniformisation » des données. L'étape suivante (troisième étape) consiste à étiqueter grammaticalement le corpus avec l'étiqueteur de Brill (Brill, 1994) et/ou le TreeTagger (Schmid, 1994). A partir du corpus étiqueté, la dernière étape extrait les unités lexicales les plus fréquentes. Notons qu'une phase d'analyse des unités lexicales obtenues peut également être ajoutée dans le processus. Cette phase est décrite dans la section 4 de cet article.

3.2 Extraction des Unités Lexicales Simples (ULS)

Nous parlons de « monde lexical » afin de désigner les mots-clés (unités lexicales simples – cf. section 3.2 ou cooccurrences lexicales – cf. section 3.3) les plus fréquents au sein d'une collection de textes. Les mondes lexicaux ont fait l'objet de différentes études, sous des appellations et des applications diverses. La terminologie est un peu floue afin de désigner ce même phénomène. Certains parlent d'isotopie sémantique (Greimas, 1986), de mots-clés thématiques (Rossignol et Sébillot, 2003), de vecteurs conceptuels (Schwab et al., 2004), de signatures thématiques (Lin et Hovy, 2000), ou encore de cartographie lexicale (Véronis, 2003). Les applications relatives à ces notions sont de divers ordre : la Traduction Automatique (Tanguy, 1997 ; 1999) ; la désambigüisation lexicale (Pichon et Sébillot (1999) ; Rossignol et Sébillot (2003)) ; l'enrichissement d'ontologies (Agirre et al., 2000, Agirre et Lopez, 2004) ; la représentation sémantique (Schwab et al., 2004) ; le résumé automatique (Riloff, 1996, Riloff et Lorenzen, 1999, Hovy et Lin, 1999, Lin et Hovy, 2000).

Notre objectif est d'extraire les mondes lexicaux de chaque corpus (Léon, 2006), en faisant émerger des similitudes et des différences lexicales. En ce qui concerne les catégories grammaticales étudiées, nous faisons l'hypothèse que les noms, les

adjectifs et les Entités Nommées (unités simples) sont lexicalement les plus pertinents.

Ces trois catégories ont été obtenues grâce à un étiquetage morpho-syntaxique proposé par le logiciel TreeTagger[3]. Les résultats obtenus en sortie se présentent sous la forme de trois colonnes, avec un terme par ligne et les informations de lemme, de forme et de catégorie grammaticale sur chaque colonne. Le tableau 1 présente un exemple de résultat obtenu avec TreeTagger.

Des	PRP :det	du
dizaines	NOM	Dizaine
de	PRP	De
messages	NOM	message
des	PRP :det	du
comités	NOM	comité
partout	ADV	partout
en	PRP	En
France	NAM	France
nous	PRO :PER	nous
sont	VER:aux:pres	être
parvenus	VER :pper	parvenir

TAB. 1 – *Exemple de résultat obtenu par TreeTagger*

Pour chaque catégorie grammaticale pertinente, nous avons extrait les *n* mots les plus fréquents de chaque corpus, ce qui a fait émerger le monde lexical classé par catégorie grammaticale.

Le tableau 2 présente les vingt noms les plus fréquents du monde lexical pour le corpus Clara. Ce dernier qui traite du féminisme est décrit de manière précise dans la section 4. Cette table montre que les mondes lexicaux sont souvent liés à une thématique donnée.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

femme	loi
an	année
droit	violence
vie	jour
homme	temps
pays	personne
monde	société
enfant	association
magazine	famille
fille	question

TAB. 2 – *Monde lexical des noms les plus fréquents du corpus « Clara »*

3.3 Extraction des Unités Lexicales Complexes (ULC)

Nous proposons par la suite d'extraire la terminologie, c'est-à-dire les cooccurrences lexicales entre deux lexèmes liés syntaxiquement et dont la fréquence est significative au sein d'un corpus. Les termes extraits appelés dans la suite des Unités Lexicales Complexes forment un monde lexical spécifique.

De multiples approches de recherche terminologique ont été développées afin d'extraire les termes pertinents à partir d'un corpus. Nous ne traiterons pas ici les approches d'aide à la structuration et au regroupement conceptuel des termes qui sont détaillés dans les travaux de (Aussenac-Gilles et Bourigault, 2003). Les méthodes d'extraction de la terminologie sont fondées sur des méthodes statistiques et/ou syntaxiques. Le système TERMINO de (David et Plante, 1990) est un outil précurseur qui s'appuie sur une analyse syntaxique afin d'extraire les termes nominaux. Cet outil effectue une analyse morphologique à base de règles, suivie de l'analyse des collocations nominales fondée sur une grammaire. Les travaux de (Smadja, 1993) (approche XTRACT) s'appuient sur une méthode statistique. XTRACT extrait, dans un premier temps, les collocations binaires situées dans une fenêtre de dix mots. Les collocations binaires sélectionnées sont celles qui dépassent d'une manière statistiquement significative la fréquence due au hasard. L'étape suivante consiste à extraire les collocations plus générales (collocations de plus de deux mots) contenant les collocations binaires trouvées à la précédente étape. ACABIT de (Daille, 1994) effectue une analyse linguistique afin de transformer les collocations nominales en termes binaires. Ces derniers sont ensuite triés selon des mesures statistiques.

Le système EXIT (Roche, 2004) consiste à extraire les termes complexes de manière itérative en utilisant des critères statistiques (mesures statistiques) et syntaxiques (patrons syntaxiques). Contrairement à ACABIT et EXIT qui sont essentiellement fondés sur des méthodes statistiques, LEXTER (Bourigault, 1993) et SYNTEX (Bourigault et Fabre, 2000) s'appuient en grande partie sur une analyse syntaxique approfondie afin d'extraire la terminologie du domaine. La méthode consiste à extraire les syntagmes nominaux maximaux. Ces syntagmes sont alors décomposés en termes de « têtes » et d' « expansions » à l'aide de règles grammaticales. Les termes sont alors proposés sous forme de réseau organisé en fonction de critères syntaxiques.

Dans notre étude, nous allons appliquer le système EXIT qui a une méthode mixte (syntaxique et statistique) afin d'extraire la terminologie nominale de base[4]. Nous nous intéressons à trois patrons morpho-syntaxiques : Nom-Adjectif, Adjectif-Nom et Nom-Préposition-Nom. Les corpus ont été étiquetés avec l'étiqueteur Brill (nécessaire pour l'utilisation des paramètres par défaut du logiciel EXIT) afin d'extraire les unités respectant ces patrons.

Le tableau 3 présente un exemple d'Unités Lexicales Complexes obtenues pour le patron Nom-Adjectif à partir du corpus Clara. Les Unités Lexicales Complexes obtenues sont analysées dans la section suivante.

mutilations génitales	communauté internationale
gynécologie médicale	volonté politique
temps partiel	concept rétrograde
junte militaire	scène slam
commission paritaire	planning familial
parlement européen	comités locaux
état civil	opinion publique
droits humains	condition féminine
acteurs sociaux	justice sociale
prisonniers politiques	journée internationale

TAB. 3 – *Unités Lexicales Complexes (Nom-Adjectif) du corpus « Clara ».*

[4] Notons que nous n'appliquerons pas le processus itératif d'EXIT.

4 Comparaison des mondes lexicaux des différentes périodes

Dans cette section, nous avons recours à une comparaison entre les mondes lexicaux d'Unités Simples et Complexes des deux périodes. Nous avons extrait les *n* premières unités, pour chaque catégorie et chaque corpus et nous avons évalué le pourcentage d'unités communes. Ainsi, un tel pourcentage s'appuie à la fois sur une notion quantitative (sélection des unités les plus fréquentes) tout en considérant leur présence (respectivement absence) lors de la comparaison des mondes lexicaux deux à deux.

Nos résultats sont analysés en deux temps. D'une part, nous avons comparé automatiquement le nombre d'unités communes. D'autre part, nous avons analysé manuellement la pertinence des unités non communes.

4.1 Description des corpus

Les corpus étudiés concernent deux revues traitant de la même thématique, le féminisme, sur deux époques distinctes. La première revue, « l'Egyptienne », est une revue mensuelle francophone diffusée en Egypte, datant des années 1930, traitant de sujets autour du féminisme tels que la politique, la sociologie, l'art, etc. Nous avons réuni 11 numéros de cette revue, qui nous ont été accessibles grâce au Centre d'Etudes Alexandrines. Ces numéros couvrent les dates de septembre 1925 jusqu'à mars 1930. Ils représentent environ 250 000 mots.

Cette revue a été numérisée par le Centre d'Etudes Alexandrines, à l'aide d'outils efficaces pour la Reconnaissance Optique de Caractères (ROC) (par exemple un appareil de prise de vues 'Phase One' qui permet de scanner de grands formats avec une forte précision pour ensuite pouvoir procéder à une reconnaissance de caractères avec une grande fiabilité). La reconnaissance optique de caractères permet ensuite de traduire des images de textes imprimés ou dactylographiés en fichiers de texte. Les fichiers obtenus par le Centre d'Etudes Alexandrines sont au format PDF. Nous les avons convertis en mode texte grâce à un logiciel de conversion de fichiers PDF en TXT[5]. Cette conversion a posé quelques difficultés liées à l'exploitation automatique de corpus « océrisés » (altération de certaines chaînes de caractères) et nous avons procédé à une phase manuelle de nettoyage.

La deuxième revue, « Clara », est une revue sur le féminisme dont les archives sont en ligne[6]. Nous avons réuni les archives disponibles, traitant de diverses thématiques telles que le racisme, l'Europe, la violence, etc. Les documents étant plus courts que pour la revue « l'Egyptienne », nous avons réuni 84 archives, qui se présentent sous la forme de dossiers traitant d'un sujet d'actualité donné. Les périodes

[5] http://www.simpopdf.com/pdf-to-text.html
[6] http://clara-magazine.fr/

s'étendent de septembre 2006 jusqu'au mois de janvier 2010. Le nombre de mots est d'environ 100 000.

4.2 Résultats

4.2.1 Analyse quantitative des unités

Le tableau 4 montre une différence entre les mondes lexicaux des unités complexes et ceux des unités simples : les mondes lexicaux des unités complexes sont relativement éloignés contrairement à ceux des unités simples. Le fait que les unités simples, et plus particulièrement les noms et adjectifs, soient très proches confirme que nous étudions une même thématique (par exemple, les unités simples *femme*, *droit*, *travail*, *enfant*, *famille*, significativement pertinentes pour notre thématique sont communes aux deux corpus).

Cependant, les Entités Nommées (EN) qui sont souvent liées à une époque restent naturellement assez spécifiques. Il en va de même pour l'analyse contrastive des Unités Lexicales Complexes, qui met en exergue des préoccupations et des spécificités d'une époque. Ceci explique donc les pourcentages d'unités complexes communes très faibles pour les unités complexes (cf tableau 4).

Ces unités seront évaluées dans la section suivante qui propose une analyse contrastive à partir des unités non communes entre les deux corpus (« Clara » et « l'Egyptienne »). Ceci nous permettra de vérifier si de telles unités sont liées au domaine du féminisme en mettant en relief les unités propres à une époque.

n	**Unités lexicales complexes (ULC)**			**Unités lexicales simples**		
	Nom-Adj	**Adj-Nom**	**Nom-prép-Nom**	**Adj**	**Nom**	**EN**
10	0%	50%	0%	70%	70%	40%
50	8%	24%	0%	64%	42%	24%
100	5%	15%	2%	50%	41%	31%
200	2.5%	9.5%	1,5%	50%	43%	42%

TAB. 4 – *Pourcentage d'unités communes parmi les n premières unités les plus fréquentes extraites (« Clara » et « l'Egyptienne »).*

Notons que les résultats sur les unités lexicales complexes Adjectif-Nom et Nom-Préposition-Nom sont à nuancer car le nombre d'occurrences de toutes les unités est moindre (égal à 1). Dans ce cas, le classement par nombre d'occurrences n'est pas un critère adapté car l'ordre proposé devient en fait aléatoire. Si nous considérons les

200 premières unités Adjectif-Nom et Nom-Préposition-Nom, nous n'obtenons aucun terme commun. Si nous sommes en condition idéale (tous les termes communs extraits avec notre système placés en tête), le pourcentage de termes communs sur les 200 premiers termes des unités Adjectif-Nom et Nom-Préposition-Nom est respectivement de 5,5% (11 termes en commun sur 200) et de 0%. Ces proportions restent très faibles et confirment que les unités lexicales complexes des deux corpus sont très différentes.

Dans la section suivante, nous verrons l'analyse qualitative de ces résultats.

4.2.2 Analyse qualitative des unités

Le tableau 5 montre que le corpus « Clara » possède une proportion d'unités pertinentes spécifiques (c.-à-d. non communes) liées au féminisme plus importante comparativement au corpus « l'Egyptienne » (49% et 62% pour les Noms et Noms-Adjectifs du corpus « Clara » vs. 37% et 40% pour le corpus « l'Egyptienne »). Une unité est jugée en relation avec le féminisme si elle traite de luttes féministes mais également, de manière plus large, des différents problèmes spécifiquement rencontrées par les femmes de chaque époque, d'où l'étude qualitative des unités non communes. Par exemple, sur la base du corpus l' « Egyptienne », le terme *mouvement féministe* a bien sûr été jugé pertinent contrairement au terme *terre cuite* lui aussi extrait automatiquement à partir de ce même corpus. Ce dernier terme traite d'une thématique spécifique d'une époque mais sans aucun lien avec le concept du féminisme traité dans cet article.

Le tableau 5 montre également que les unités complexes sont plus largement liées au féminisme que les unités simples pour chaque corpus. Ces résultats confirment que les unités complexes sont, naturellement, plus spécifiques à une thématique et à une époque que les unités simples, comme par exemple *planning familial*, *orientation sexuelle*, *harcèlement sexuel*, etc. pour le corpus « Clara » et *mortalité infantile*, *fièvre puerpérale*, etc. pour le corpus l'« Egyptienne ». Notons que de nombreuses préoccupations des années 1930 sont relatives à la santé des femmes ce que le corpus l'« Egyptienne » met parfaitement en exergue avec les termes *mortalité infantile* et *fièvre puerpérale* (maladie infectieuse survenant après un accouchement ou une fausse couche).

Pertinence des Unités	Nom		Nom-Adjectif	
	Clara	Egyptienne	Clara	Egyptienne
Unités pertinentes (liées au féminisme)	49%	37%	62%	40%
Unités pertinentes mais trop générales (non liées au féminisme)	44%	59%	38%	59%
Unités non pertinentes	7%	3%	0%	1%

TAB. 5 – *Pourcentage des 100 premières unités non communes évaluées manuellement : unités de type Nom (ULS) et Nom-Adjectif (ULC) extraites à partir des deux corpus (« Clara » et « l'Egyptienne »).*

5 Conclusion

Nous avons présenté une analyse lexicale diachronique entre deux corpus traitant de la même thématique, sur une période différente. L'analyse lexicale s'est appuyée sur l'émergence et la comparaison des mondes lexicaux des deux corpus. Ces mondes lexicaux, unités lexicales les plus fréquentes, mettent en valeur la thématique et l'univers lexical d'un corpus. La comparaison de ces mondes lexicaux permet d'une part de confirmer la similarité thématique entre deux corpus et d'autre part de mettre en valeur les spécificités de chaque corpus, qu'il s'agisse de sous-thématiques spécialisées ou de divergences diachroniques. Ce type d'étude pourrait être utile dans d'autres contextes applicatifs, comme par exemple celui de la désambigüisation lexicale ou de la Traduction Automatique.

Dans nos futurs travaux, nous souhaitons améliorer la phase de nettoyage des données afin d'obtenir des corpus moins bruités. Nous souhaitons par ailleurs effectuer un classement des unités fondé sur des mesures statistiques plus adaptées que la fréquence. En effet, cette dernière ne prend pas en compte la répartition des unités dans les différents documents d'un même corpus contrairement à d'autres critères tels que les mesures TF-IDF et/ou OKAPI.

Sur la base d'étude en corpus, de nombreux travaux de fouille de textes (Grouin *et al.* 2011) étudient les spécificités lexicales voire syntaxiques des données afin de mettre en exergue des variations diatopiques (selon une positon géographique), diastratiques (selon la dimension sociale ou démographique) ou diachroniques comme nous le montrons dans cet article. Nous avons concentré notre étude sur la présence/absence d'unités entre deux types de corpus afin de mettre en avant les spécifici-

tés lexicales d'une époque donnée. Cependant, il pourrait être intéressant de mener une étude plus fine afin de mesurer les différentes variations lexicales ou sémantiques liées au féminisme (mots, syntagmes ou concepts présents sous des formes différentes selon les époques).

Références

AGIRRE, E., OLATZ, A., HOVY, E., MARTINEZ, D. (2000). Enriching very large ontologies using the WWW. Ontology Construction of the European Conference of AI (ECAI), Berlin, Allemagne.

AGIRRE, E., LOPEZ, O. (2004). Publicly available topic signatures for all wordnet nominal senses. Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal.

AUSSENAC-GILLES N., BOURIGAULT D. (2003), Construction d'ontologies à partir de textes. Actes de TALN, Volume 2, p27–47.

BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de Grammaires 25, p131–151.

BOURIGAULT D. (1993), Analyse syntaxique locale pour le repérage de termes complexes dans un texte. TAL, 34(2), p105–118.

BRILL E. (1994), Some advances in transformation-based part of speech tagging. In AAAI, Vol. 1, pp. 722–727.

DAILLE B. (1994), Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. Ph.D. thesis, Univ. Paris 7.

DAVID S., PLANTE P. (1990), De la nécéssité d'une approche morpho syntaxique dans l'analyse de textes. In Intelligence Artificielle et Sciences Cognitives au Quebec, Volume 3, pp. 140–154.

GROUIN C., FOREST D., PAROUBEK P., ZWEIGENBAUM P. (2011). Présentation et résultats du défi fouille de texte DEFT2011 - Quand un article de presse a t-il été écrit ? À quel article scientifique correspond ce résumé ? Actes du septième défi fouille de texte (DEFT), pp. 3-14.

HOVY, E., LIN C. Y. (1997). Automated Text Summarization in SUMMARIST. Workshop on Intelligent Scalable Text Summarization, Madrid, Espagne.

LEON S. (2006), Acquisition automatique de traductions de termes complexes par comparaison de « mondes lexicaux » sur le Web. Actes de RECITAL, p700-709.

LIN C.-Y., HOVY E. (2000), The Automated Acquisition of Topic Signatures for Text Summarization. Actes de COLING.

PICHON R., SÉBILLOT P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience. Actes de Traitement Automatique des Langues Naturelles (TALN).

RILOFF E., LORENZEN J. (1998). Extraction-Based Text Categorization: Generating Domain-Specific Role Relationships Automatically, Natural Language Information Retrieval, p167-196

ROCHE M. (2004), Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes. Ph.D. thesis, Univ. Paris 11.

ROSSIGNOL M., SEBILLOT P. (2003), Extraction statistique sur corpus de classes de mots-clés thématiques. TAL, 44(3), p217-246.

SMADJA F. (1993), Retrieving collocations from text : Xtract, Computational Linguistics, Vol. 19, pp. 143-177.

SCHWAB S., LAFOURCADE M., PRINCE V. (2004), Hypothèses pour la construction et l'exploitation conjointer d'une base lexicale sémantique basée sur les vecteurs conceptuels. Actes des JADT, Louvain-la-Neuve, Belgique.

TANGUY, L. (1997). Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration informatique d'un modèle de la sémantique interprétative. Thèse de doctorat, Ecole Nationale Supérieure des Télécommunication de Bretagne.

TANGUY L. (1999). Isotopies sémantiques pour la vérification de traduction. Traitement Automatique des Langues Naturelles (TALN).

SCHMID H. (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the Int. Conf. on New Methods in Language Processing, p44-49.

VERONIS J. (2003), Cartographie lexicale pour la recherche d'information. Actes de TALN, Batz-sur-Mer, France

Remerciements

Un grand merci à Louis-Jean Calvet (Université de Provence), au Centre d'Etudes Alexandrines et particulièrement à Jean-Yves Empereur, pour avoir initié ce projet et mis à notre disposition toutes les ressources numérisées de la presse francophone.

Summary

This paper presents a diachronic comparative analysis between two corpora dealing with the domain of feminism, on two different periods. Lexical analysis is based on the acquistion of "lexical worlds" (i.e. simple and complex lexical units significantly frequent) related with both corpora and on a comparative analysis of these worlds. The results show that the simple lexical units are very similar between both corpora that deal with the same topic, while the complex lexical units are significantly different, because they are more specialized to a sub-topic and a period.